Grounded Representations for Sensory-Motor Learning: A Linguistic Approach

Gutemberg Guerra-Filho (GUERRA@Cs.Umd.Edu)
Department of Computer Science, A. V. Williams Bldg.
University of Maryland, College Park, MD 20742 USA

Yiannis Aloimonos (YIANNIS@Cfar.Umd.Edu)
Center for Automation Research, A. V. Williams Bldg.
University of Maryland, College Park, MD 20742 USA

Abstract

We have empirically discovered that the space of human actions has a linguistic structure. This is a sensory-motor space consisting of the evolution of the joint angles of the human body in movement. The space of human activity has its own phonemes, morphemes, words (nouns, verbs, adjectives, adverbs), and sentences formed by syntax. This has a number of implications for the grounding problem and cognition in general. We present a Human Activity Language (HAL) for symbolic manipulation of visual and motor information. The embodiment of the language serves as the interface between visual perception and the motor subsystem. The visuo-motor language is defined using a linguistic approach. In phonology, we define basic atomic segments that are used to compose human activity. In morphology, we study how visuo-motor phonemes are combined to form strings representing human activity and to generate a higher-level morphological grammar. In syntax, we present a model for visuo-motor sentence construction where the subject corresponds to the active joints (noun) modified by a posture (adjective). A verbal phrase involves the representation of the human activity (verb) and timing coordination among different joints (adverb).

Keywords: grounding problem, sensory-motor intelligence.

Introduction

Natural intelligent systems perceive events occurring in the environment, reason about what is happening, and act accordingly. This process involves mapping observed motor sequences onto a vocabulary of actions. This vocabulary represents motor patterns performed previously and stored according to some knowledge representation.

An artificial intelligence with commensurate abilities may require a symbolic structure for reasoning about human activities. However, the semantic interpretation of a symbolic representation system, such as natural language, cannot be based only on meaningless arbitrary symbols.

The symbol grounding problem, Harnard (1990), addresses this semantic gap and suggests that the primitives of a formal symbolic system should be associated with grounded representation connected to physical experience in the world.

A grounded representation is a sensory-motor projection of objects and events to which elementary symbols refer. In this paper, we concentrate in events associated with human activities. This way, a sensory-motor projection consists in the translation from a non-symbolic analog representation of human activities in the world to a grounded non-arbitrary symbolic representation according to invariant features, which allow cognitive tasks such as recognition.

The sensory-motor projection of primitive words leads to language grounding. Language grounding for verbs has been addressed by Siskind (2001) and Bailey at al. (1997) from the perspective of perception and action, respectively.

Biological evidence, such as the functionality of Broca's region in the brain (Nishitani at al. 2005) and the mirror neurons theory (Gallese at al. 1996), suggests that perception and action share the same symbolic structure as a knowledge that provides common ground for recognition and motor planning.

Furthermore, spoken language and visible movement use a similar cognitive substrate based on the embodiment of grammatical processing. With evidence that language is grounded on the motor system (Glenberg and Kaschak 2002), we propose a visuo-motor language as a grounded representation for sensory-motor learning.

Our visuo-motor language is called Human Activity Language (HAL). HAL is specified in a linguistic approach, where phonetics, morphology and, syntax are defined. The linguistics framework is used to represent human movement with a symbolic system. However, the symbols have a non-arbitrary mapping to the sensory-motor primitives. A linguistic approach benefits from the theory of Automatic Speech Recognition and Natural Language Processing.

In kinesiology and movement analysis, the symbolic representation materializes the concept of motor programs and enables the identification of common motor subprograms used in different activities.

In Computer Vision, a visuo-motor language would allow the visual parsing of human movement which may be used in action recognition and video annotation to extract symbolic descriptions from real-world data.

The visuo-motor language may also help humanoid robots to generalize the planning and control of motor activities while using a vocabulary of human actions. In Computer Graphics, this language could lead to a different approach in animation programming.

We propose HAL as a representation which allows the use of parsing and symbolic reasoning about information

related to human activities. This representation is compact and, consequently, computationally efficient.

Sensory-Motor Embodiment

Mirror neurons are brain cells which activate when a monkey performs a specific action and also fire when the monkey observes the same action (Gallese at al. 1996). There is evidence that mirror neurons exist in human brains. This suggests a common representation for perceptual and motor information.

A visuo-motor language plays a central role in supporting activity understanding as a common representation for visual and motor information. A perception system takes the visual input and extracts higher-level representations for human actions. These representations are parsed and possibly matched to visuo-motor programs by the recognition process. If the action vocabulary does not contain the observed action, no matching is found and learning occurs through imitation. The imitation process searches for a physically feasible plan to execute the observed unknown action in the action subsystem.

Visual Representations

Visual and motor aspects of human activity are abstracted to a common ground through embodiment which is, ultimately, the consideration of the human body into the modeling process. In this paper, we focus in the discovery of a common embodied symbolic language.

One instance of the visual perception process is achieved by a Motion Capture (MoCap) system. We captured videos featuring 90 different human activities and the corresponding three-dimensional reconstruction for trajectories of body parts was found using our own MoCap system, Guerra-Filho (2005). Given this three-dimensional reconstruction, joint angles were computed to describe human movement.

Motor Representations

Muscles are stimulated by electrical impulses (action potentials) that travel from a nerve to a muscle. The nerve is activated when a threshold current is achieved. Each nerve action potential activates the muscle propagating another action potential into the muscle fibers to cause contraction.

All basic moves a human body can perform result from single muscle activations. Different muscles collaborate to perform some specific anatomic action on a particular body part. Anatomic actions are the most basic movements that are visible and, hence, they are a starting point for the cyclic cognitive process between visual and motor representations. In general, anatomic actions can be divided into flexion/extension, abduction/adduction, and rotations. An anatomic action performed by a specific joint and occurring in a particular anatomic plane (traverse, frontal, sagital) corresponds to a degree of freedom (DOF) in a human body.

A joint angle time-varying function for all DOFs represents a human action. This approach to action definition is used in this paper for the derivation of a grounded symbolic representation: the Human Activity Language.

Kinetology

A first process in our linguistic methodology is to find structure in human movement through basic units akin to phonemes in spoken language. Kinetology is the study of the basic atoms of human movement, including the motion representation, segmentation, and rules modeling and constraining these atoms.

We propose the concept of a kinetological system and five principles on which such a system should be based: compactness, view-invariance, reproducibility, selectivity, and reconstructivity.

We use human walking gait to illustrate and evaluate our kinetological system. Human walking is a well constrained action, involves several coordinated body parts, and it has been extensively studied. Actual movement data of human walking gait is analyzed.

Compactness and View-Invariance

The compactness principle is related to describing a human activity with the least possible number of atoms. We define a state according to the sign of derivatives of the original 3D motion. The derivatives used in our segmentation are velocity and acceleration.

A view-invariant representation provides the same 2D projected description of an intrinsically 3D action captured from different viewpoints. In order to evaluate view-invariance, a circular surrounding configuration of viewpoints is used.

A compactness/view-invariance (CVI) graph for a DOF shows the states associated with the movement according to two dimensions: time and space (see Fig. 1). For each time instant (horizontal axis) and for each viewpoint in the configuration of viewpoints (vertical axis), the movement state is associated with a representative color.

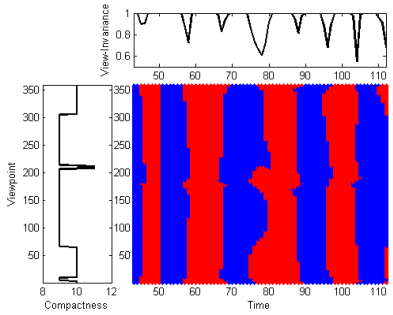


Fig. 1: Compactness/View-Invariance Graph.

The compactness measurement consists in the number of segments when the movement varies with time. For each viewpoint, the compactness measurement is plotted on the left side of the CVI graph.

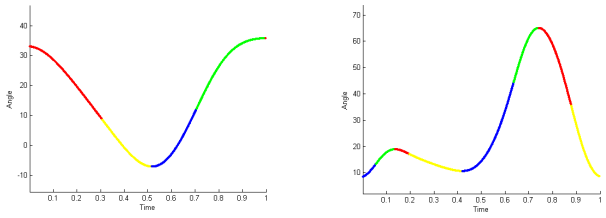
The view-invariance measurement concerns the fraction of the most representative state among all states considering all viewpoints at a single instant in time. For each time instant, the view-invariance measure is computed and plotted on the top of the CVI graph.

The view-invariance measure is affected by some uncertainty at the borders of the segments and at degenerate viewpoints (see Fig. 1). The border effect shows that movement segments are not completely stable during the temporal transition between segments. The degenerate viewpoints are special cases of frontal views where the sides of a joint angle tend to be aligned.

Repeatability

An important requirement for a kinetological system is the ability to represent actions exactly in the same way even facing inter-personal or intra-personal variability. A kinetological system is repeatable when the same symbolic representation is associated with the same action performed by different subjects.

Each segment corresponds to an atom α, where α is a symbol associated with the segment's state. The atomic symbols (**R**, **Y**, **B**, **G**), called kinetemes, are the phonemes of our kinetological system. The symbol **R** is assigned to negative velocity and negative acceleration segments; the symbol **Y** is assigned to negative velocity and positive acceleration segments; the symbol **B** is assigned to positive velocity and positive acceleration segments; and the symbol **G** is assigned to positive velocity and negative acceleration segments (see Fig. 2).



(a) Hip Flexion/Extension (b) Knee Flexion/Extension

Fig. 2: Symbolic representation of joint angle functions.

A repeatability measure is computed for each joint angle considering a gait database. The repeatability measure of a joint angle is the fraction of the most representative symbolic description among all descriptions for the database files. A very high repeatability measure means that symbolic descriptions match among different gait files and, consequently, the kinetological system is repeatable.

The repeatability measure is very high for the joint angles which play a primary role in the walking action (see Fig. 3). Using our kinetological system, six joint angles obtained very high repeatability: pelvic obliquity, hip flexion-extension, hip abduction-adduction, knee flexion-extension, foot rotation, and foot progression. These variables seem to be the most related to the movement of walking forward. Other joint angles obtained only a high repeatability measure which is interpreted as a secondary role in the action: pelvic tilt and ankle dorsi-plantar flexion. The remaining joint angles had a poor repeatability rate and seem not to be correlated to the action purpose but, probably, to its stability.

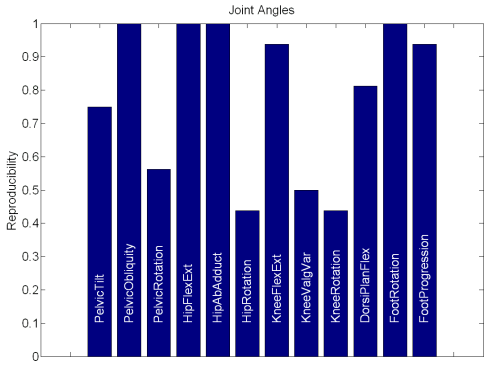


Fig. 3: Repeatability measure for 12 DOFs during gait.

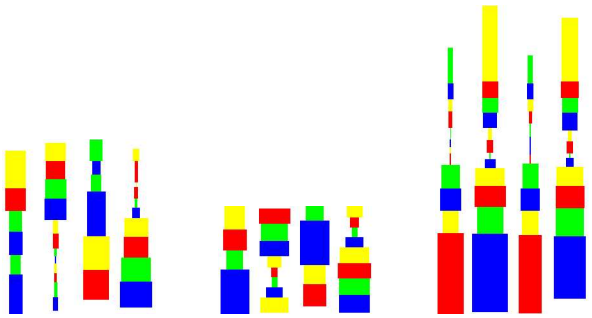
Our kinetological system performance on the repeatability measure for all the joint angles shows that the system is repeatable for the DOFs intrinsically related to the action. Further, the system is useful in the identification (unsupervised learning) of the variables playing primary roles in the activity. The identification of the intrinsic variables of an action is a byproduct of the repeatability requirement of a kinetological system.

Selectivity

The selectivity principle concerns the ability to discern between distinct actions. In terms of representation, this principle requires a different structure to represent different actions. We compare the compact representation of several different actions and verify whether their structures are dissimilar.

Our representation has a qualitative aspect, the state of each segment, and a quantitative aspect corresponding to the time length and angular displacement of each segment. The qualitative aspect is depicted with colors, while the quantitative aspect is represented by the line segment length and thickness for time length and displacement, respectively.

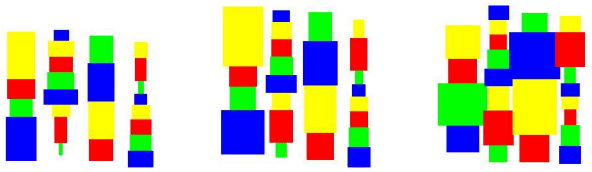
The selectivity property is demonstrated in our representation using a set of three actions performed by the same individual. Four joint angles are considered: left and right hip flexion-extension, left and right knee flexion-extension. The three examined actions are walk, run, and jump (see Fig. 4).



(a) Walk (b) Run (c) Jump

Fig. 4: Compact representations of five different actions.

The three different actions are clearly represented by different structures with respect to the qualitative aspect. However, manner variations are only different in the quantitative aspect. We investigate the quantitative aspect of manner variations of the walk action (see Fig. 5).



(a) Walk (b) Walk with stride (c) Walk with exaggerated stride

Fig. 5: Four manner variations of the walk action.

Each manner variation has a total of 24 segments for the four joint angles considered. For a pair of variations, we compute a dissimilarity vector, where each element corresponds to the difference between the quantitative aspects of the associated segments in the two variations. The least median dissimilarity of the manner variations was 12%. This way, even for the same action, the representation has enough dissimilarity to select between different manner variations.

Reconstructivity

Reconstructivity is associated with the ability to reconstruct the original movement signal up to an approximation factor from a compact representation. Once the movement signal is segmented and converted into a symbolic representation, this compact description is only useful if we are able to recover the original joint angle function or an approximation.

In order to use the sequence of kinetemes for reconstruction, we consider one segment at a time and concentrate on the state transitions between two consecutive segments. Based on a transition, we determine constraints about the derivatives at border points of the segment. Therefore, we investigate the possible state transitions that are feasible in our kinetological system.

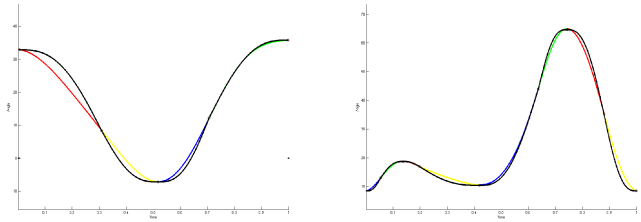
For this discussion about reconstructivity, let's assume that the signs of velocity and acceleration don't change simultaneously. This way, each segment can have two possible next neighbor segments. However, the transition **B** $\rightarrow$ **Y** (**R** $\rightarrow$ **G**) is impossible, since velocity cannot become negative (positive) with a positive (negative) acceleration.

From the kinetological rules, there are eight possible sequences of three consecutive segments. Each possible sequence corresponds to two equations and two inequality constraints associated with first and second derivatives at border points. Other two inequalities come from the derivatives at interior points of the segment.

A simple model for the joint angle function during a segment is a polynomial. However, low degree polynomials don't satisfy the constraints originated from the possible sequences of kinetemes. For example, a cubic function has a linear second derivative which is impossible for sequences where the second derivative assumes zero value at the borders and non-zero values at interior points (*e.g.* GBG). The least degree polynomial satisfying the constraints imposed by all possible sequences of kinetemes is a fourth degree polynomial. This way, the reconstruction process consists in finding the five parameters defining this polynomial with the two associated equations for the particular sequence of kinetemes.

Using a fourth degree polynomial, additional constraints are required to reconstruct the joint angle function corresponding to a segment. We introduce two more equations coming from the joint angle values at the two border points. These values are the time length and the angular displacement of each segment. Angular displacement is the absolute difference between initial and final joint angle of a segment.

With four equations, a linear system is solved up to one variable. This last free variable is constrained by four inequalities and it can be determined using some criteria such as jerk (third derivative) minimization (see Fig. 6).



(a) Hip Flexion/Extension (b) Knee Flexion/Extension

Fig. 6: Reconstruction of joint angle functions.

Morphology

Given the kinetological representation for an activity praxicon (lexicon of movement), a hierarchical organization is derived in the form of morphological grammars for the activity lexicon. The kinetological strings for each activity represent the lowest level in the morphological grammar (see Fig. 7). Each segment i is represent by a state α_i and a displacement Δ_i. The

displacement Δ of a segment is a quantization of the absolute difference between the initial and final joint angles of a segment.

jog := B0 G0 R2 Y3 B3 V0 G3 R0 Y0

jump := G0 R0 Y0 B2 G1 V0 B5 G3 R0 Y0 R10 Y5 B0 G0 R0 Y0 B0 G1 B0 G0 R0 Y1 B0 V0

run := B4 G7 R1 Y2 B1 G0 R3 Y5 V0 Y0

scuff := **R0 V0 Y0 V1 Y0 V0 R0 V0 Y0** V0 R0 V0 Y0 B0 V0 G0 V1 G0 V0 G0 □0 V0 □0

stomp := G0 R0 Y0 V0 R1 V0 Y2 V1 R0 V0 Y1 V0 Y0 B0 G0 V0 B1 V0 G1 V0 B1 V0 G2 R2 Y

swagger := Y0 V0 R1 V0 Y0 V0 R0 V0 Y0 V0 □0 R0 V0 Y0 V0 R0 V0 Y0 B2 G2 V0 B0 V0 G0

tiptoe := B0 G1 V1 B0 V0 G1 □0 B0 V0 G0 **R0 V0 Y0 V0 Y0 V0 R0 V0 Y0** B0 V0 G0 R0 V0 Y0

toe := R0 V1 R0 V0 Y0 V0 R0 Y0 B0 G1 V2 G1 R0 V0 Y0 B0 G0

troop := R0 Y0 B0 G0 R1 V3 Y3 V1 Y0 B5 G3 R0 Y0 B2 G3 R2 V0 Y2 B0 G0

walk := **R0 V0 Y0 V1 Y0 V0 R0 V0 Y0** B2 G2 V0 B0 V0 G0 R0 V0 Y0 □0 R0 Y1 V0 R0 V0 Y1

Fig. 7: Morphological grammar at the lowest level.

The generation of a higher-level morphological grammar involves finding common digrams in different activities of the lexicon. Our algorithm finds the most frequent pair $\alpha_i\Delta_i$ $\alpha_j\Delta_j$ of consecutive atoms in the current grammar. A new grammar rule $L_n := \alpha_i\Delta_i\ \alpha_j\Delta_j$ is then created. A higher-level grammar is generated using the new rule. Each occurrence of the pair of atoms $\alpha_i\Delta_i\ \alpha_j\Delta_j$ in the current grammar is replaced by a non-terminal L_n. This process is repeated until the most frequent pair in the current grammar has less than two occurrences and, consequently, the highest level of the grammar is reached.

The highest level of the grammar contains the lexical units of HAL. The sub-string alphabets embed the structure that allows the identification of roots, prefixes, and suffixes in the lexical units. Furthermore, this structure implies relations that give rise to a hierarchical organization.

Syntax

In a sentence, a noun represents the subjects performing an activity or objects receiving an activity. A noun in a HAL sentence corresponds to the body parts active during the execution of a human activity and to the possible objects involved passively in the action.

The initial posture in an activity is analogous to an adjective in a HAL sentence which further describes (modifies) the active joints (noun) in the sentence. The HAL adjective is represented by a string of integers considering only the active joints in the activity. Each element in this string corresponds to the quantized initial angle of an active joint (see Fig. 8).

	jog	jump	run	scuff	stomp	swagger	tiptoe	toe	troop	walk
R_Hip	4	4	2	4	5	2	1	3	4	2
R_Knee	7	7	4	8	6	8	8	9	7	9
R_Ankle	8	5	4	5	4	3	0	3	4	4
L_Hip	3	3	4	2	3	1	3	2	2	2
L_Knee	1	9	7	9	7	8	9	7	6	7
L_Ankle	5	4	7	4	3	0	4	0	5	3

Fig. 8: HAL adjectives for an experimental lexicon.

The sentence verb represents the changes each active joint experiences during the action execution. The representation for a HAL verb was discussed previously. However, further description is required to deal with coordination among different joints.

A coordinated segment is a time interval delimited by events representing local minima and maxima in the joint angle function for any of the active joints. These events occur in between specific atomic pairs (YΔ BΔ and GΔ RΔ) and, consequently, may be computed from the HAL verb strings (see Fig. 9).

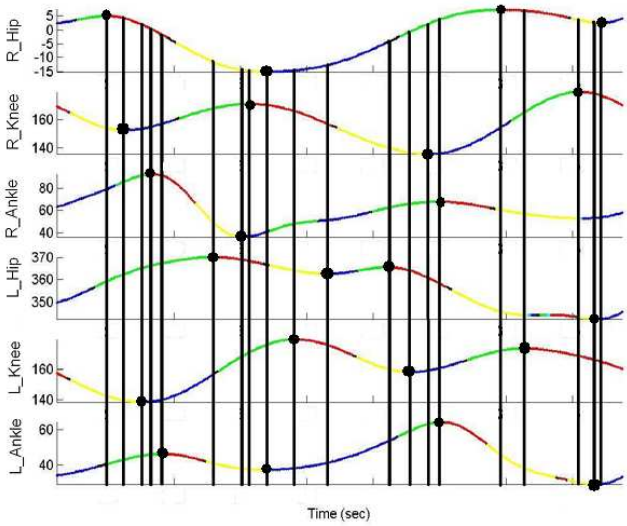


Fig. 9: Coordinative segments for the jog action.

A HAL adverb is a string of multiplicative constants modeling the variation in the execution time and displacement of each coordinated segment. A HAL adverb is appended to a verb in such a way that each value in the adverb string corresponds to a coordinated segment in the verb.

The Subject-Verb-Object (SVO) pattern of syntax is a reflection of the patterns of cause and effect. An action is represented by a word that has the structure of a sentence: the agent or subject is a set of active body parts; the action or verb is the motion of those parts. In many such words, the action is transitive and involves an object or another patient body part.

A HAL sentence S := NP VP consists of noun phrase (noun + adjective) and verbal phrase (verb + adverb), where NP := N Adj and VP := V Adv (see Fig. 10). The organization of human movement is simultaneous and sequential. This way, the basic HAL syntax expands to orthogonal axis based on joint (parallel syntax) and time (sequential syntax) structure. The parallel syntax concerns simultaneous activities represented by parallel sentences $S_{t,\ j}$ and $S_{t,\ j+1}$ and constrains the respective nouns to be different: $N_{t,\ j} \neq N_{t,\ j+1}$. This constraint states that

simultaneous movement must be performed by different body parts. An example of two parallel sentences is an action with walk and wave simultaneously.

The temporal sequential combination of action sentences ($S_{t,j}\ S_{t+1,j}$) must obey the cause and effect rule. The HAL noun phrase (body parts) must experience the verb cause (motion) and the joint configuration effect must lead to a posture corresponding to the noun phrase of the next sentence. Considering noun phrases as points and verb phrases as vectors in the same space, the cause and effect rule becomes $NP_{t,j} + VP_{t,j} = NP_{t+1,j}$ (see Fig. 10). The cause and effect rule is physically consistent and embeds the ordering concept of syntax.

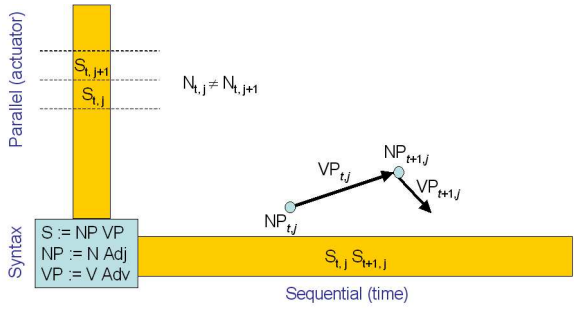


Fig. 10: Parallel and sequential syntax.

The lexical units are arranged into sequences to form sentences. A sentence is a sequence of actions that achieve some purpose. In written language, sentences are delimited by punctuation. Analogously, the action language delimits sentences using motionless actions such as stop, still, and freeze, for example. In general, a conjunctive action is performed between two actions, where a conjunctive action is any preparatory movement that leads to an initial position required by the lexical unit.

Conclusions

In this paper, we presented a visuo-motor language, named Human Activity Language – HAL, which is the basic kernel for symbolic manipulation of visual and motor information of human activity. The visual information is processed in a perception system which translates a visual representation of action into an embodied representation matched to our visuo-motor language. Our instance of the visual process is a Motion Capture system which reconstructs human movement from images.

The embodiment is an important characteristic of the language serving as interface between visual perception and the motor planning towards action execution. The visuo-motor language is defined using a linguistic approach by specifying phonology, morphology, and syntax. This methodology is one possible way to construct a visuo-motor language which may build upon the established results of speech recognition and natural language processing.

In phonology, we presented a suggestion for the basic atomic segments that are used to compose human activity. Segments are characterized according to the sign of the first and second angular derivatives of joints.

In morphology, we studied how HAL phonemes were combined to form strings representing human activity. Basically, we explored common substrings to generate a higher-level morphological grammar which is more compact and suggests the existence of lexical units working as visuo-motor subprograms.

In syntax, we presented a model for visuo-motor sentence construction where the subject in a sentence corresponds to the active joints (noun) modified by a posture (adjective). HAL verbs represent the changes each active joint experiences during an activity execution. Coordination among different joints is specified by timing the atomic segments and appending elastic discrete values (adverb) to coordinated segments. The adverb is used to adjust and modify the action execution.

For future work, we plan to investigate other morphological grammar generation algorithms and evaluate these algorithms according to the compactness of the language. The phonology and morphology of nouns, adjectives, and adverbs are issues that deserve more attention.

References

Bailey D., Chang N., Feldman J., & Narayanan S. (1997). Extending Embodied lexical development. *Proceedings of the 20th Annual Meeting of the Cognitive Science Society.*

Browman, C. & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V. Fromkin (Ed.), *Phonetic linguistics.* New York: Academic Press.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, CA., & Rizzolatti, G. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. *Journal of Cognitive Neuroscience*, 16, 114-126.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593-609.

Glenberg, A. & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3), 558-565.

Guerra-Filho G. (2005). Optical Motion Capture: Theory and Implementation. *Brazilian Computing Society Revista de Informática Teórica e Aplicada*, 12(2), 61-89.

Harnard S. (1990). The symbol grounding problem. *Physica*, D42, 335-346.

Nishitani, N., Schurmann, M., Amunts, K., & Hari, R. (2005). Broca's region: from action to language. *Physiology*, 20, 60-69.

Siskind J. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15, 31-90.